OIST OKINAWA INSTITUTE OF SCIENCE AND TECHNOLOGY GRADUATE UNIVERSITY

Transcription factor binding modeling
Rotation report

Student Name: Darya Stepanenko
Student ID: 1701029
Date of Submission: September 5, 2018
Unit: Biological complexity Unit, Simone Pigolotti

I. AIMS

Tissues are distinguished by gene expression patterns, specific for every tissue type. Gene expression regulates by a wide range of mechanisms that are used by cells to increase or decrease the production of a specific gene. One of the mechanisms is regulation by Transcription factors (TFs). Transcription factors bound at a remote site helps RNA polymerase bind to the promoter and start transcribing.

Although the regulation by TFs has a crucial role on a cell's faith, the whole process of recognition and binding is poorly understood. The diffusion of the TF toward the DNA and the recognition of the binding site happen faster than a diffusion allows. Therefore other models are suggested, such as two-modes diffusion achieved by three-dimensional diffusion and one-dimensional sliding along the DNA sequence [1]. This computational study aims to understand the specification of the sliding process.

II. BACKGROUND

The TF slides along the DNA chain with weak specificity until finds specific $5-10$ nucleotides long sequence, called motif, and stopes the movement by strong attachment to the motif. However experimental data shows that only a fraction of motifs are actually bound by TFs. The sliding along the surrounding of a motif may influence the binding process by switching the weak binding to the strong in necessary conditions [2].

The influence of the DNA content around the motif on the binding between DNA and transcription factors have been already studied by different groups. For example, it has been shown that Homeodomain TFs prefer binding to AT rich regions [3], while such TF families as C2H2 and ETS prefer GC rich content [2]. Also the group in Helsinki [4] has analyzed 830 binding profiles of human TFs using high-throughput systematic evolution of ligands by exponential enrichment SELEX [5] and ChIP sequencing, describing 239 distinctly different binding specificities. It proved that the environment of the bound motifs demonstrates unique sequence compositions for different TF families, DNA shape features, and overall high similarity to the binding motif. Although the above results made the picture more clear it this still a question how the content of the DNA chain in non motif areas influences the switching between weak and strong specificity.

III. MODEL

DNA structure in eukaryotic cells is extremely complex for understanding due to the influence of a chromosomal packaging. Therefore the technique based on artificially produced sequences, called systematic evolution of ligands by exponential enrichment SELEX [6], that allows the measurement of the binding intensities of one transcription factor to numerous synthetic double-stranded DNA sequences in a single experiment.

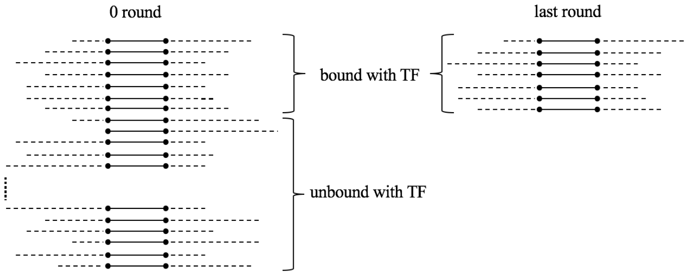

FIG. 1: The scheme of a collection of a set of bound sequences and a set of unbound sequences from SELEX.

Systematic evolution of ligands by exponential enrichment (SELEX) is a method to enrich small number of bound DNAs from a random sequence pool by PCR amplification. It provides a powerful way to determine in vitro the binding specificities of transcription factors.

SELEX is characterized by number of rounds [see Fig.1]. In the first round transcription factor binds to a mixture of random set of DNA sequences around 30 bp in length, containing recognized motif and flanked by primers that allow PCR amplification. Motifs bound by a transcription factor are then PCR amplified and used as an input in a further DNA-binding reaction, and so on. After multiple rounds of selection (around 3-4), the pool is sequenced to determine which sequences are preferentially bound by the transcription factor.

Based on SELEX data the differences of the DNA context of bound and unbound sequences could be defined. The position specific scoring matrix based on frequencies $P_{\alpha,j}$ of letter α detection at each position j around the motif in the sequence determines how often the nucleotide $\alpha = \{A, T, G, C\}$ appears at a chosen position j among all sequences, that contain the motif [see Eq.1].

$$P_{\alpha,j} = \frac{N_{\alpha,j}}{N_{all\_bound}}, \tag{1}$$

where $P_{\alpha,j}$ is the probability of the nucleotide $\alpha = \{A, T, G, C\}$ to appear at a chosen position j among all bound sequences N_{all_bound}.

IV. METHOD

Bioinformatic pipeline consists of some key steps:

- Motif recognition;
- Alignment;
- Analysis of the alignment.

Motif recognition

Experimental SELEX data is collected from the open database European Nucleotide Archive under the number ERP001824, cited in [4]. For bound sequences the final SELEX selection round is used, whereas for unbound sequences initial pool of sequences and round$_{last-1}$ are considered.
The target motif for each transcription factor is assumed to be known and collected from the open database JASPAR, that has approximately 500 TFs and its motifs.

For each transcription factor the binding motifs is considered with allowance of:

- one mutation, that means Hamming distance is equal to 1;
- no mutation, Hamming distance is equal to 0.

In each round of SELEX PCR is performed, that lead to the big amount of repeats among the data. To get read of repeats the command fastx_collapser from the command line tool, called fastx_toolkit, is used. That allows to leave only uniq sequences for next step of analysis. For example from roughly 130 000 seqs in last round only 25 000 uniq seqs are left, that also decreases the time for the next alignment step.

Allingnment

To get the distribution of letters around the motif the alignment is needed [see Fig.2]. The alignment was made with the command line tool called Clustalw2, with a high gap penalty parameter in order to forbid gaps.

```
------------GCCGTAACCGCAAACCGCATGATCNGCACG--...---
------------GCCGTAACCGCAAACCGCAGGATCTGCACG--...---
------------GCCGTAAACGCAAACCGCAAGATCNGCACG--...---
------------GCCGTAACCGCAAACCGCAAGATCCGCACG--...---
------------AGGGAAACCGCAAAGCCGCAAGTACGCACG--...---
------TACTACGTATCACCGCAAAACCGCAAGTGG--------...---
-------------ACGTATCCGCAAACCGCAAGGGGCGCGTAA-...---
-----------ACGGGTAACCGCAAACCGCNCGTGGCGCAT---...---
-----------ACGGGTAACCGCAAACCGCGCGTGNCGCAT---...---
-----------ACGGGTAACCGCAAACCGCACGTGCCGCAT---...---
-----------ACGGGTAACCGCAAACCGCACGTGACGCAT---...---
-----------CGGCGAAACCGCAAACCGCAGTCGCCTTGG---...---
-----------CGGCGAAACCGCAAACCGCCGTCGNCTTGG---...---
--------------CAATACCGCAAACCGCACCTCGTGTTGGAG...---
------------GCCGTAACCGCAAACCGCAGCTTGTGTGGC--...---
------------GCCGTAACCGCAAACCGCAGCTTGNGTGGC--...---
----------AGGGCGTAACCGCAAACCGCGCCTTGTATG----...---
----------AGGGCGTAACCGCAAACCGGGCCTNGTATT----...---
...
-------TCAGACCGCAAAACGTAACAGCCANATTAG-------...---
```

FIG. 2: Sample of the alignment of 15789 sequences.

Analysis

Based on the position specific scoring matrix [Eq.1] of aligned sequences [Fig.2] the content of nucleotides of bound sequences (round$_{last}$) and all sequences with the binding motif (round$_0$ and round$_{last-1}$) is revealed. Two visualization approaches are considered:

- an alignment representation through logo [7], where the motif is visible [see Fig.3],
- A+T nucleotides frequency as a function of the position in a sequence.

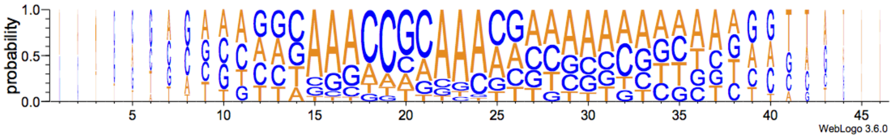

FIG. 3: Representation of the alignment with logo.

V. RESULTS

The pipeline allows to observe the content of nucleotides in sequences, which contain the transcription factor binding motif, and distinguish the specificity of the content, that is in charge of switching the weak sliding to strong binding. The built pipeline was tested on one of transcription factors, called RUNX3.

Fixed non mutable motif

The assumption that the transcription factor recognizes the exact motif without allowance for any mutations is made. It turned that in the initial pool sequences with fixed binding motif, claimed in the database, were not present. Moreover in round$_{last}$ and round$_{last-1}$ the patterns of A+T frequency exhibit very similar tendencies [see Fig.4]. Based on that it worth to consider that:

- transcription factors are tolerant to mutations in the binding regions;
- round$_{last-1}$ and round$_{last}$ could not be compared for unbound and bound sequences as they show very similar binding patterns.

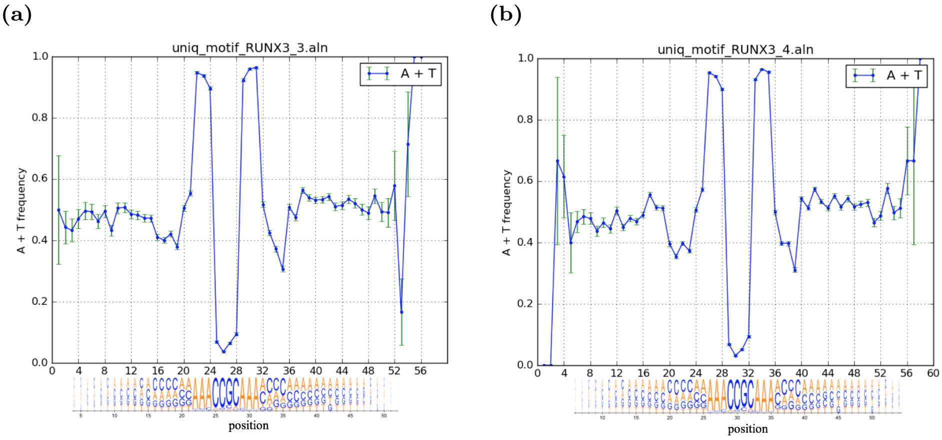

FIG. 4: Frequency of A + T nucleotides among sequences that contain the motif in the experiment **(a)**SELEX round$_{last-1}$**(b)** SELEX round$_{last}$

Motif with one mutation

As the level of mutation in eukaryotic cells is around 100 mutations per genome, the tolerance to mutations of the transcription factor is reasonably acceptable. Following the above idea the system with the allowance of one mutation in the motif recognition site is considered.

As shown in Fig.5a, in the initial pool of sequences the A+T concentration around the binding side fluctuates around 50%. While bound sequences, round$_{last}$ and round$_{last-1}$, exhibit the increase of A+T content along some positions before the motif in the sequence [see Fig.5b and Fig.5c]. Although no statistical tests were done to prove the relevance of the hypothesis that the initial pool and the bound sequences are different, the error bars show that the increase in frequency has a relevant meaning even despite the noise.

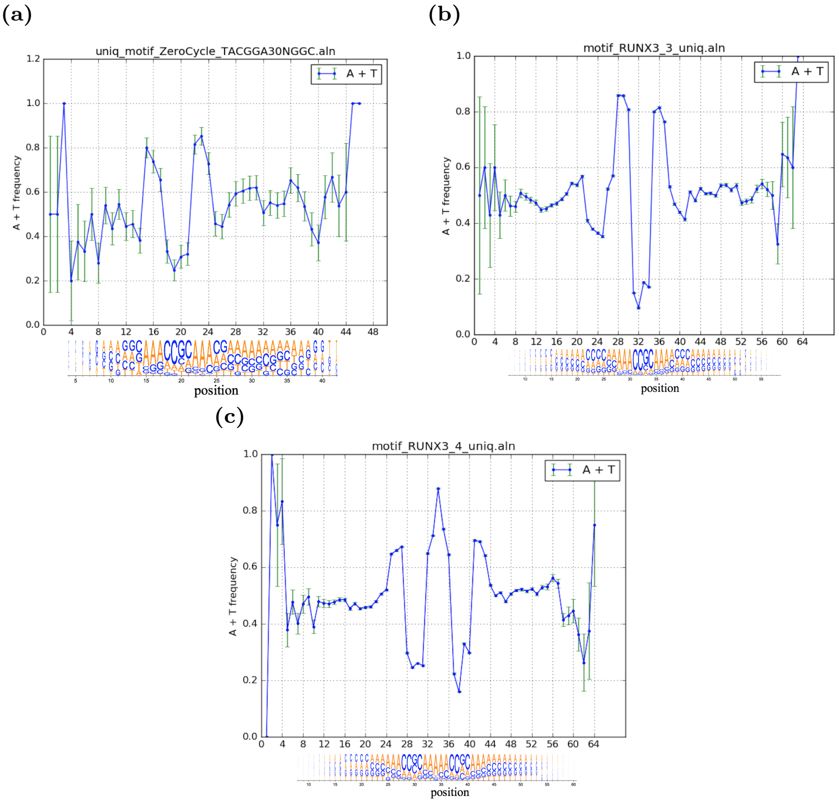

FIG. 5: Frequency of A + T nucleotides among sequences that contain the motif and its analogues with one mutation in the experiment **(a)** initial pool of sequences **(b)** SELEX round$_{last-1}$ **(c)** SELEX round$_{last}$

VI. CONCLUSION

This work was aimed as a first step toward the understanding of the two-modes diffusion of the transcription factor in eukaryotes. We built a computational pipeline, that allows to analyze the specificity in nucleotide content around the transcription factor's binding site.

Additional plans aim to:

- apply the pipeline to different transcription factors families;
- build a physical model of the sliding process based on Boltzmann energy distribution revealed from position specific scoring matrix;
- study the mechanism of regulation by transcription factors as the whole process: three-dimensional diffusion search, one-dimensional sliding, recognition and binding, unbinding.

The code was written in python and shell script. Such command line programs as clustalw2 2.1, weblogo and fastx_toolkit also were used. The code is available at OIST Dropbox and could be shared if necessary.

[1] Massimo Cencini and Simone Pigolotti. Energetic funnel facilitates facilitated diffusion. *Nucleic Acids Res*, 46(2):558–567, Jan 2018.

[2] Iris Dror, Tamar Golan, Carmit Levy, Remo Rohs, and Yael Mandel-Gutfreund. A widespread role of the motif environment in transcription factor binding across diverse protein families. *Genome Res*, 25(9):1268–80, Sep 2015.

[3] Nicoletta Bobola and Samir Merabet. Homeodomain proteins in action: similar dna binding preferences, highly variable connectivity. *Curr Opin Genet Dev*, 43:1–8, Apr 2017.

[4] Arttu Jolma, Jian Yan, Thomas Whitington, Jarkko Toivonen, Kazuhiro R Nitta, Pasi Rastas, Ekaterina Morgunova, Martin Enge, Mikko Taipale, Gonghong Wei, Kimmo Palin, Juan M Vaquerizas, Renaud Vincentelli, Nicholas M Luscombe, Timothy R Hughes, Patrick Lemaire, Esko Ukkonen, Teemu Kivioja, and Jussi Taipale. Dna-binding specificities of human transcription factors. *Cell*, 152(1-2):327–39, Jan 2013.

[5] D Irvine, C Tuerk, and L Gold. Selexion. systematic evolution of ligands by exponential enrichment with integrated optimization by non-linear analysis. *J Mol Biol*, 222(3):739–61, Dec 1991.

[6] Arttu Jolma and Jussi Taipale. Methods for analysis of transcription factor dna-binding specificity in vitro. *Subcell Biochem*, 52:155–73, 2011.

[7] Gavin E Crooks, Gary Hon, John-Marc Chandonia, and Steven E Brenner. Weblogo: a sequence logo generator. *Genome Res*, 14(6):1188–90, Jun 2004.